This is the Author Accepted version of the paper L. Candela, D. Castelli, P. Pagano "The D4Science Experience on Virtual Research Environments Development" in Computing in Science & Engineering, 2023, doi: 10.1109/MCSE.2023.3290433

CiSE SI – Science Gateways: Accelerating Research and Education
Editor: Patrick Diehl, patrickdiehl@lsu.edu – Rafael Ferreira da Silva, silvarf@ornl.gov

The D4Science Experience on Virtual Research Environments Development

L. Candela
Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

D. Castelli
Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

P. Pagano
Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

***Abstract*—Today, complex research challenges, often based on the analysis of a large amount of data, require multidisciplinary collaboration and appropriate communication and sharing of data, processes and outcomes. Technologies and large-scale infrastructures provide stakeholders with computing capacity and data services to perform unprecedented levels of data-driven scientific activities. This opens the way to science gateways and virtual research environments supporting researchers in scientific and educational activities. This article describes our extensive experience with the Virtual Research Environments (VRE) operated by the D4Science infrastructure. It presents how this infrastructure supports their development, their basic functionalities and how they are easily customised to serve the needs of specific user communities. It also describes how they are used in real contexts. The article concludes by reporting how VREs are now progressively used as valuable instruments to support open science and how this role might become more relevant in the future.**

VIRTUAL RESEARCH ENVIRONMENTS (VREs) and Science Gateways are solutions aiming at providing a designated community with an online research platform catering to integrated access to *resources* (e.g., computing, software, data, instruments) of interest for the community [1], [2], [3]. Several approaches and technologies were proposed to implement these typologies of solutions [4].

This article presents our experience with the D4Science infrastructure and VREs development in the last 18 years [5], [6]. This experience started with the DILIGENT project "*to create an advanced test-bed allowing members of dynamic virtual e-Science organisations to access shared knowledge and to collaborate in a secure, co-*

ordinated, dynamic and cost-effective way" [7]. The main idea of this project was to develop a cyber-infrastructure to deliver *VREs as-a-Service*. Since then, the development of this infrastructure, later named D4Science, as of the VREs features, continued with the support received via many EU Commission-funded projects and other funding initiatives.[1] In all these projects, VREs were used to serve domain-specific user communities and use cases. Our research approach to improve the VRE solution has always been translational [8], i.e., the application cases have been intrinsically bound into the research and development project timeline rather than being an optional and separate activity. By doing this, the feedback collected from the VREs users has been fed back into the VREs construction project itself. This partnership approach has been essential to enable effective and efficient use of the VREs in real-world contexts. The experience so far has also enabled us to understand the more frequent needs, recognise clusters of co-occurring requirements, and identify success factors and additional desired functionality.

We describe the D4Science approach and its support for creating and operating multiple VREs. Then, we discuss the landscape of the VREs created, followed by the most relevant lessons learned. Finally, an outlook on the role VREs will play in the future concludes the paper.

OVERVIEW

D4Science-based VREs are web-based, community-oriented, collaborative, user-friendly, open-science-enabler working environments for scientists and practitioners willing to work together to perform a specific (research) task. From the end-user perspective, each VRE manifests in a unifying web application hosted in a web gateway (and a set of application programming interfaces (APIs)) comprising several components made available by portlets organised in custom pages and menu items running in a plain web browser. Every component aims to provide VRE users with facilities implemented by relying on one or more services, possibly provisioned by diverse providers. Every VRE plays the role of a community-specific web application giving seamless access to the datasets and services of interest for the designated community while hiding the diversities originating from various resource providers. Among the components each VRE offers, some basic ones are enacting VRE users to perform their tasks collaboratively, namely: (*a*) a workspace component to organise and share any digital artefact of interest; (*b*) a social networking component to communicate with co-workers by posts and replies; (*c*) a data analytics platform to share and execute analytics methods; (*d*) a catalogue component to document and publish any digital artefact worth sharing.

The D4Science infrastructure [5], [6] is at the heart of the solution for creating, maintaining and operating VREs as-a-Service. It provides the enabling services and the necessary resources to implement them. Figure 1 depicts the service-oriented view of the D4Science infrastructure architecture (for details, refer to previous works [5], [6]). Services are conceptually organised into three groups: (*i*) front end components, the D4Science part with which the user interacts directly; (*ii*) back end components, the D4Science part implementing the business logic of the system; and (*iii*) provided resources, the D4Science part providing front-end components and back-end components with resources (computing, storage, data, software) to use.

The D4Science front end manifests into a series of Liferay[2] portal instances and several REST APIs for accessing the services serving a specific VRE. These instances are either a replica of the portal or the service implementing one of the APIs. Instances are made available by a (high availability) proxy implementing load balancing policies, that is, distributing the calls to the existing service instances to balance the load. The Liferay portal is equipped with portlets specifically conceived to give access to functionalities offered by one or more D4Science back end components. A new specific site (Liferay concept) is created in the portal instance to host and realise the VRE front end of each VRE.

The D4Science back end components are services (often frameworks on their own) organised in four main areas: (*i*) core services support-

[1] https://en.wikipedia.org/wiki/D4Science

[2] www.liferay.com

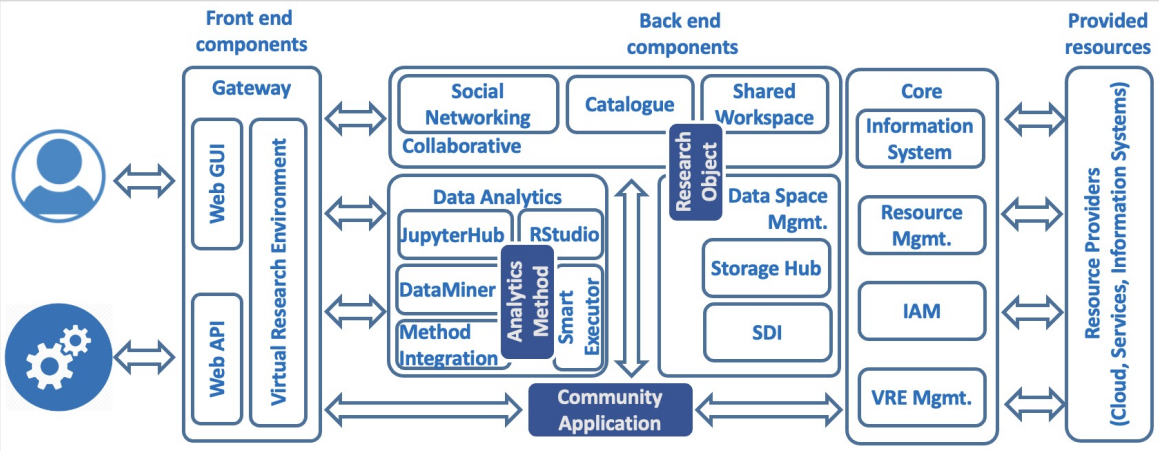

Figure 1. The D4Science Service-oriented Architecture

ing VREs management, resources management, authentication and authorisation; (*ii*) data space management services supporting the storage and management of various typologies of data, including files stored in diverse storage systems and geospatial data managed by a spatial data infrastructure (SDI) exploiting an array of orchestrated GeoNetworks[3], THREDDS data servers[4], and GeoServers[5]; (*iii*) data analytics services supporting several options for data analytics, including the DataMiner proprietary platform with its integration facility [9], JupyterHub[6], and a cluster of RStudio[7] instances; (*iv*) collaborative services implementing facilities enacting the collaboration among the members of a VRE, for example, by supporting communication and sharing.

The D4Science provided resources include the D4Science distributed computing infrastructure. This distributed computing infrastructure is spread across four main sites (CNR-Pisa, GARR-Catania, GARR-Naples, GARR-Palermo)[8], geographically distributed, and managed across different administrative domains. It also exploits resources from the EGI federation[9] and resources operated by the Copernicus DIAS service[10].

[3]https://geonetwork-opensource.org/
[4]https://www.unidata.ucar.edu/software/tds/
[5]https://geoserver.org/
[6]https://jupyter.org/
[7]http://rstudio.org/
[8]The National Research Council of Italy (CNR) www.cnr.it and the GARR Consortium www.garr.it, the Italian National Research and Education Network, currently host the D4Science sites.
[9]www.egi.eu
[10]www.copernicus.eu/en/access-data/dias

IMPACT

The D4Science infrastructure has supported the delivery of VREs for very diverse communities and usage scenarios. The creation of these VREs has been a continuous process. Some VREs were created for communities continuing to use and maintain them, while others were dismissed upon completing the activity that had motivated their deployment.

At the time of writing this manuscript (June 2023), D4Science operates 20 thematic gateways[11] and more than 190 Virtual Research Environments (with others to come). These environments support the activities and tasks of diverse communities of practice ranging from marine science to social science, humanities, agri-food, health, and geothermal science.

The D4Science userbase counts over 21,000 users from almost all over the world. Fig. 2 displays the growth of the user base in the last five years. The growing trend is clear and constant.

Fig. 3 displays the number of working sessions performed by D4Science users in the last five years. In total, more than 440,000 working sessions have been executed, with an average of circa 6,800 working sessions per month. Over 1.2 billion analytics tasks were completed, with an average of circa 18.5 million tasks per month.

The so far created VREs can be classified according to the major requirements they were called to address: *Communication and sharing*, *Analytics*, *Education*, *Publishing*, and *Service*

[11]https://services.d4science.org/thematic-gateways

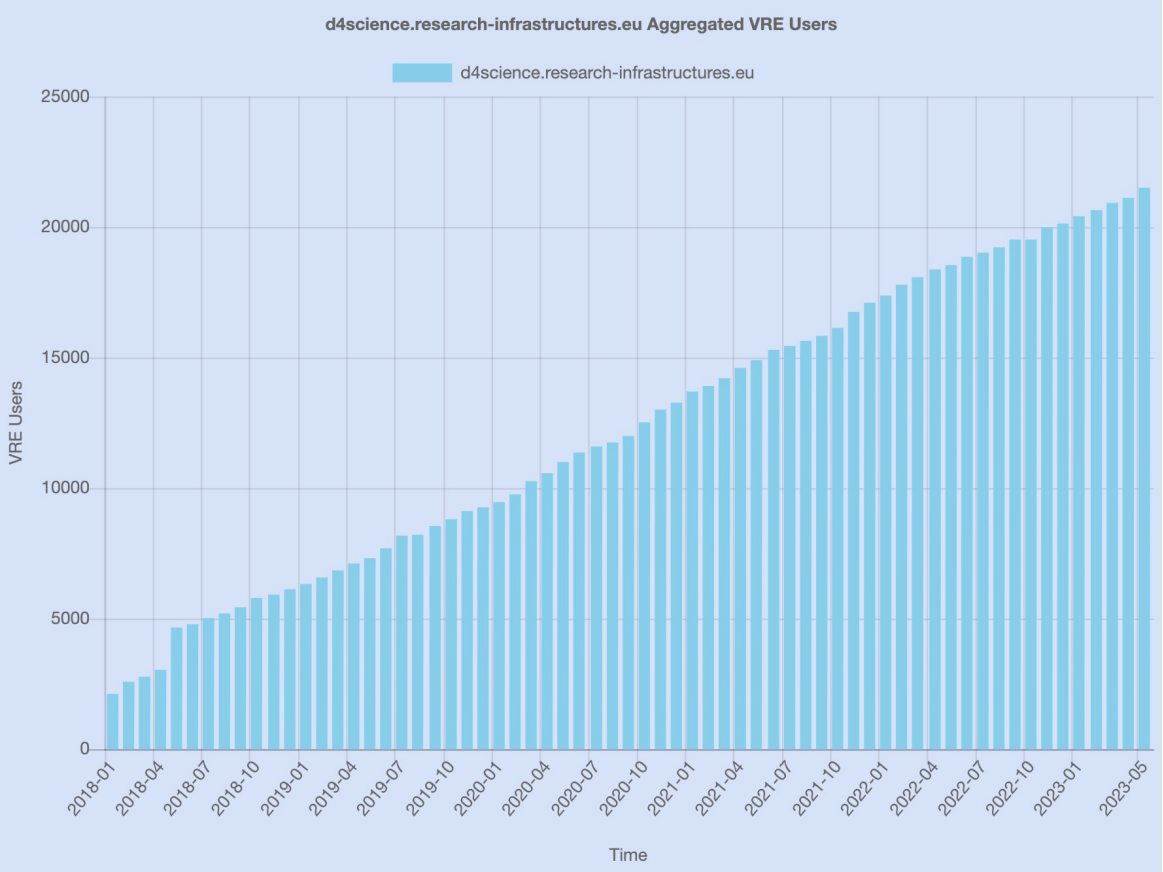

Figure 2. D4Science Userbase January 2018-May 2023

provisioning. These typologies of VREs are not disjoint from the functional point of view; rather, they are characterised by a primary demand that implies a peculiar exploitation of the offered functionalities. We briefly describe each of these classes and some examples of them.

Communication and sharing

These are the basic requirements for any collaborative activity. VREs typically designed to serve these requirements are mainly demanded by teams collaborating to manage shared projects with working groups often spread across diverse institutions and regions. They offer basic working environments, providing their designated team with a workspace for sharing artefacts (mainly documents) and a social networking area for discussing. D4Science hosts many of these VREs dedicated to EU, national and thematic project teams, e.g., the European Research Infrastructure for Heritage Science (E-RIHS), the European Open Science Cloud working groups, and the Italian Oceanographic Commission (the National Coordination Body of the Intergovernmental Oceanographic Commission of UNESCO).[12]

Analytics

The primary demand of communities performing data-centred research tasks is to have user-friendly solutions for data analytics capable of relying on a powerful and distributed computing infrastructure and hiding any technical complexity. The requested environments are also expected to support the sharing of the developed analytics solutions and the reproducibility of the processes executed. The D4Science VREs offer three possible solutions made available in various settings for responding to the analytics requirements: the DataMiner platform with the dashboard for executing shared analytics pipelines and reproducing experiments, the JupyterLab for implementing and executing analytics tasks by notebooks, and

[12] https://services.d4science.org/explore

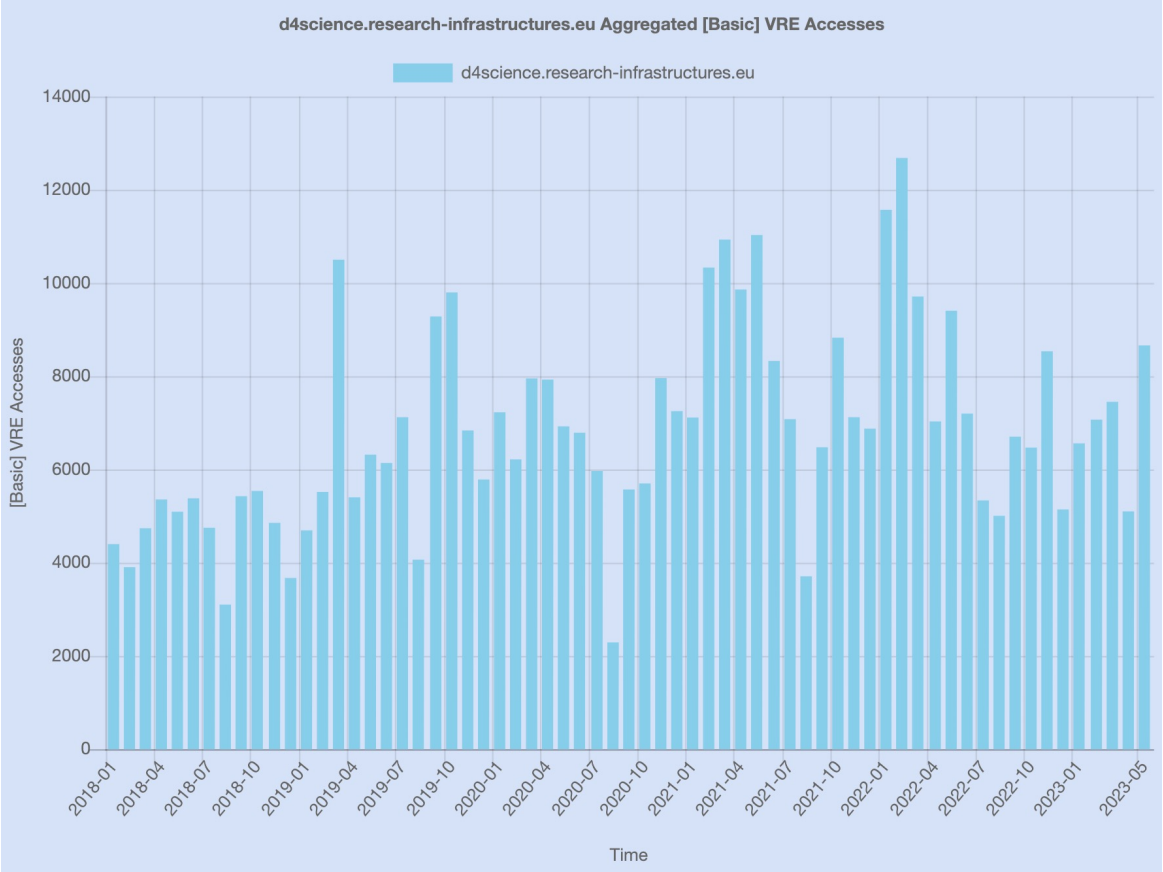

Figure 3. D4Science Working Sessions January 2018-May 2023

RStudio for R-based analytics. These analytics solutions are available in many VREs, including a couple of openly available and community-agnostic environments: AnalyticsLab[13] (offering all three analytics solutions) and RStudioLab[14] (offering RStudio only). By using these openly available environments, diverse analytics tasks were accomplished, including computing a high-resolution global-scale model for COVID-19 infection rate, batch video processing for fish size detection, and computing geographical suitability maps for geothermal power plants.

Education and Training

Over time, we have also seen a great demand for virtual environments dedicated to education and training. There is often a gap in making the necessary technological support available to scientists, trainees and students in specialised university courses and training events and workshops. This is particularly true when dealing with interdisciplinary scenarios and contexts. VREs facilitate the setting up and delivery of training courses in a cost-effective way and make available collaboration and integrated access to potentially unlimited digital research resources, as well as cross-disciplinary and cross-community tools and services. Examples of this type of VREs are: (*i*) BiodiversityLab[15], an environment designed to support University courses by providing a collection of applications that allow scholars to perform complete experiments in the ecological domain (e.g., inspect species maps and produce new ones using either an expert system or a machine learning model, perform analysis of climatic changes and their effects on species distribution, discover Taxa names, cluster occurrence data, and estimate similarities among habitats); (*ii*) SDG Indicator 14.4.1[16], a learning environment for FAO training

[13] https://services.d4science.org/web/analyticslab
[14] https://services.d4science.org/web/rstudiolab
[15] https://services.d4science.org/web/biodiversitylab
[16] https://i-marine.d4science.org/web/sdg-indicator14.4.1

activity on monitoring, analysing and understanding of stocks and fisheries data under the control of a dedicated trainer. Users/trainees can find or upload relevant datasets for exercises related to fisheries data analysis and rely on an interactive R Shiny application[17] for interactive editing of algorithm parameters and results visualisation.

Publishing

Another primary demand in developing data-driven approaches is to collaboratively collect, curate and make publicly findable and accessible data products. VREs responding to this need usually have communication and sharing facilities and a dedicated catalogue. The typologies of catalogue items are fully defined by the designated community in terms of types of products (e.g., datasets, services, training material, research objects) and metadata formats characterising each item. Examples of VREs deployed to meet this primary demand are (*i*) the EOSC-Pillar Training and Support Catalogue [10] used by an editorial team to collect and publish research data management training material collaboratively, (*ii*) the GRSF VRE [18] designed to handle the information on fish stock monitoring that countries perform directly or through the Regional Fishery Bodies (RFBs). Once harmonised, this information is made accessible through a unique catalogue, the Global Record of Stocks and Fisheries, valuable for a large variety of actors, including RFBs and their member states, the seafood industry, national agencies, researchers and officers, and NGOs; (*iii*) the FMJ Lab [11] supporting the publishing of the executable versions of the analytics pipelines presented in papers published by the Food and Ecological Systems Modelling Journal[19].

Service Provisioning

A recurring request that we receive is also to facilitate the development and delivery of one or more community-specific services to a designated community. This typology of VRE serves two classes of users: the service providers and the end users. Service providers develop and operate value-added services by exploiting the general-purpose services offered by D4Science. For instance, this is the case of the Marine Environmental Indicators VLab[20]. An innovative and task-specific web application was developed by designing and implementing only the front-end part and outsourcing the computing part to the DataMiner platform. The time and effort needed to deliver the new facility were reduced with respect to from-scratch development thanks to the functionality the VRE offers as well as to the hosting environment offered by the underlying infrastructure. Another example is represented by the family of entity linking tools made available by the TagME VRE[21]. The management of users, the hosting of the services and the computing capacity were borrowed from the VRE the underlying infrastructure by obtaining a higher availability with respect to that guaranteed before moving to D4Science.

LESSON LEARNED

The growing number of users and cases supported, their diversity, as well as the long-lived and expanded exploitation of D4Science-based VREs from organisations like FAO, ESFRI Research Infrastructures, EU and national projects demonstrate that the overall D4Science solution is effective for many. Here below, we summarise the key features that, according to the feedback received, make it a viable and successful solution for developing Virtual Research Environments.

The ***VREs as-a-Service*** delivery approach is, to a large extent, the most relevant feature of the D4Science solution for most user communities. Most do not have the necessary resources and personnel to deploy, host and maintain such environments. Often they are also looking for solutions to help them to minimise the "time-to-market", i.e., the time in which they can start using the VRE to support their specific activities. D4Science implements a VRE distribution model in which it hosts the whole application and provides it to users over the internet as a service [9]. The advantage of this design choice is that the actual management of the IT solution is in the hands of expert operators who manage it by providing reliable services, leveraging economies of scale, and using elastic approaches to scale. A new VRE is created by using a

[17] https://shiny.posit.co/
[18] https://i-marine.d4science.org/web/grsf
[19] https://fesmj.pensoft.net/
[20] https://blue-cloud.d4science.org/web/marineenvironmentalindicators
[21] https://sobigdata.d4science.org/web/tagme/

wizard to select the VRE's functional constituents among those available. The software components' deployment and configuration implementing the selected functionalities are completely automatic. It leads to a new and ready-to-use VRE made available through one of the gateways operated by D4Science.

The ***system of systems*** approach [12] is paramount to promote the establishment of synergies with several service providers and to enlarge the capacity and service offering exploitable when creating and operating VREs. In fact, D4Science was designed to conceptually play the role of a central-hub offering seamless access to its own resources (datasets, services, computing and storage capacity) as well as to services and computing capacity offered by other infrastructures and service providers. All the resources aggregated by the federated service providers are registered into a unifying information system, monitored, and exploited on demand to contribute to the creation and operation of the various VREs.

Catering for ***co-creation*** is paramount to guarantee community uptake and the incremental evolution of the VRE to meet the designated community changing needs. Communities of practice have evolving needs and often refine their requirements when using the provided working environments and services. VREs cannot be static environments; they must evolve, making available new tools and datasets to meet emerging needs. D4Science VREs support integration patterns [6] to complement the offering and bring new resources into VREs by facilitating the integration of community-specific existing applications, analytics methods and workflows, datasets and other resources for discovery and access. This co-creation mechanism counts on the presence of a working version of the VRE where community resources are "hot-plugged" without stopping or shutting down the environment.

Open science is here to stay, yet it must be supported by an open access approach "as early as convenient". This approach is progressively affirming as the new norm in science. It implies collaboration and sharing, reproducibility and transparency to a wider and greater extent possible. Scientific communities willing to operate in line with this approach have found in the D4Science VREs concrete support for flexibly meeting these properties and implementing open science practices with the needed shades of openness. D4Science VREs are equipped with basic services supporting collaboration and co-operation among its users, namely: (*i*) a shared workspace to store and organise any version of a research artefact; (*ii*) a social networking area to have discussions on any topic (including working version and released artefacts) and be informed on happenings; (*iii*) data analytics solutions to execute processing tasks either natively provided by VRE users or borrowed from other VREs; and (*iv*) a catalogue-based publishing platform to make the existence of a certain artefact public and disseminated. These facilities are at the fingertips of VRE users. They also continuously and transparently capture research activities, authors and contributors, as well as every by-product resulting from every phase of a typical research lifecycle, thus offering a solid base for addressing open science practices like, for example, reproducibility, research assessment, communication and collaboration, and transparency.[5].

CONCLUSION

This paper has presented the D4Science VREs solution, and the experience acquired so far in developing and operating a large variety of VREs addressing diverse needs.

At the current stage, we can state that this solution has many advantages, as demonstrated by its high uptake. One of the most appreciated is indeed its delivery mode. VREs as-a-Service represents, for many communities (especially communities of practice in long-tail science), the ideal solution for solving the need for their collaborative activities, especially when these are data-driven and computationally intensive and go beyond the boundaries of institutions and regions. Indeed, they largely reduce the time for a community of practice to become operational and the need for skilled personnel dedicated to technology development.

We will continue reinforcing the services offered by the underlying infrastructure to empower further the facilities that can be made available via VREs. Planned development directions include: (*i*) increase the support to scientific workflows by injecting intelligent solutions based on machine learning and recommender systems technologies;

(*ii*) enhance the attention to open science practices by improving interaction and collaboration means by supporting FAIRness-by-design and multiple forms of publication for any type of intermediate results. The aim is also to reach a point where VREs are themselves research results that are published and shared as part of the scientific communication process.

ACKNOWLEDGMENT

The experiences and developments presented in this paper are the results of a team effort done with the contribution of the D4Science team. We thank all the members of this team.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Blue Cloud project (grant agreement No. 862409), EOSC-Pillar project (grant agreement No. 857650), and SoBigData-PlusPlus (grant agreement No. 871042).

REFERENCES

1. L. Candela, D. Castelli, and P. Pagano, "Virtual Research Environments: An Overview and a Research Agenda," *Data Science Journal*, vol. 12, no. 0, pp. GRDI75–GRDI81, 2013.
2. M. Barker, S. D. Olabarriaga, N. Wilkins-Diehr, S. Gesing, D. S. Katz, S. Shahand, S. Henwood, T. Glatard, K. Jeffery, B. Corrie, A. Treloar, H. Glaves, L. Wyborn, N. P. C. Hong, and A. Costa, "The global impact of science gateways, virtual research environments and virtual laboratories," *Future Generation Computer Systems*, vol. 95, pp. 240–248, Jun. 2019.
3. P. Calyam, N. Wilkins-Diehr, M. Miller, E. H. Brookes, R. Arora, A. Chourasia, D. M. Jennewein, V. Nandigam, M. Drew LaMar, S. B. Cleveland, G. Newman, S. Wang, I. Zaslavsky, M. A. Cianfrocco, K. Ellett, D. Tarboton, K. G. Jeffery, Z. Zhao, J. González-Aranda, M. J. Perri, G. Tucker, L. Candela, T. Kiss, and S. Gesing, "Measuring success for a future vision: Defining impact in science gateways/virtual research environments," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 19, Oct. 2021.
4. M. Arezoumandan, L. Candela, D. Castelli, A. Ghannadrad, D. Mangione, and P. Pagano, "Virtual Research Environments Ethnography: a Preliminary Study," in *Proceedings of the 14th International Workshop on Science Gateways, Trento, Italy*. CEUR-WS.org, 2022.
5. M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, P. Pagano, G. Panichi, and F. Sinibaldi, "Enacting open science by D4Science," *Future Generation Computer Systems*, vol. 101, pp. 555–563, Dec. 2019.
6. M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, A. Dell'Amico, L. Frosini, L. Lelii, M. Lettere, F. Mangiacrapa, P. Pagano, G. Panichi, T. Piccioli, and F. Sinibaldi, "Virtual research environments co-creation: The D4Science experience," *Concurrency and Computation: Practice and Experience*, Mar. 2022.
7. L. Candela, F. Akal, H. Avancini, D. Castelli, L. Fusco, V. Guidetti, C. Langguth, A. Manzi, P. Pagano, H. Schuldt, M. Simi, M. Springmann, and L. Voicu, "DILIGENT: integrating digital library and Grid technologies for a new Earth observation research infrastructure," *International Journal on Digital Libraries*, vol. 7, no. 1-2, pp. 59–80, Oct. 2007.
8. M. Parashar and D. Abramson, "Translational computer science for science and engineering," *Computing in Science & Engineering*, vol. 23, no. 05, pp. 5–6, 2021.
9. M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano, G. Panichi, C. Perciante, and F. Sinibaldi, "The gCube system: Delivering Virtual Research Environments as-a-Service," *Future Generation Computer Systems*, vol. 95, pp. 445–453, Jun. 2019.
10. P. O. Garcia, L. Berberi, L. Candela, I. Van Nieuwerburgh, E. Lazzeri, and M. Czuray, "Developing the EOSC-Pillar RDM Training and Support Catalogue," in *Linking Theory and Practice of Digital Libraries*, G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, and A. Poggi, Eds. Cham: Springer International Publishing, 2022, vol. 13541, pp. 274–281, series Title: Lecture Notes in Computer Science.
11. M. Assante, A. Boizet, L. Candela, D. Castelli, R. Cirillo, G. Coro, E. Fernández, M. Filter, L. Frosini, T. Georgiev, G. Kakaletris, P. Katsivelis, R. Knapen, L. Lelii, R. M. Lokers, F. Mangiacrapa, N. Manouselis, P. Pagano, G. Panichi, L. Penev, and F. Sinibaldi, "Realizing virtual research environments for the agri-food community: The AGINFRA PLUS experience," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 19, Oct. 2021.
12. M. W. Maier, "Architecting Principles for Systems-of-Systems," *INCOSE International Symposium*, vol. 6, no. 1, pp. 565–573, Jul. 1996.

Leonardo Candela is computer science senior researcher at the National Research Council of Italy,

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo". He received his PhD in Information Science from the University of Pisa in 2006. His research interests are driven by the development of systems and services supporting research infrastructures for science by intertwining virtual research environments, data infrastructures, collaborative working environments, reference models for complex systems, information retrieval, data analytics, data publishing and innovative scholarly communication practices. His research activity is developed by closely connecting research and development. He has been and is involved in several EU-funded projects called to develop Digital Libraries & Data Infrastructures, and he is the Strategy and Portfolio Manager of the D4Science.org infrastructure. Contact him at leonardo.candela@isti.cnr.it.

Donatella Castelli is Research Director at Istituto di Scienza e Tecnologie dell'Informazione, "A. Faedo" of the National Research Council of Italy where she leads the InfraScience research Lab. Under her supervision, the InfraScience team coordinated and participated in several EU and nationally-funded projects on Digital Libraries and Research Data Infrastructures. In particular, she has been the co-ordinator of the EU projects that have developed the D4-Science infrastructure and the technical coordinator of those that have developed the OpenAIRE one. She participated in expert groups dedicated to shaping the European Open Science Cloud. She is currently the Italian member of the EU Group of National contact points for scientific Information. Her research interests include open science data infrastructures and open science scientific approaches. She authored several research papers in these fields. Contact her at donatella.castelli@isti.cnr.it.

Pasquale Pagano is Senior Researcher at Istituto di Scienza e Tecnologie dell'Informazione, "A. Faedo" (ISTI) of the National Research Council of Italy. He has a strong background and experience in models, methodologies and techniques for the design and development of distributed virtual research environments requiring the handling of heterogeneous computational and storage resources provided by Grid and Cloud based e-Infrastructures, and the management of heterogeneous data sources. He participated in the design of the most relevant distributed systems and e-Infrastructure enabling middleware developed by ISTI-CNR. He is currently the Technical Director of the D4Science Data Infrastructure. He is also the Principal Investigator of the BlueCloud2026 EU-funded project and the main contact point for other EU and Italian national projects that plan the development and operation of a variety of VREs. Contact him at pasquale.pagano@isti.cnr.it.